

**Philosophische** Fakultät III

Sprach-, Literatur- und Kulturwissenschaften

Institut für Information und Medien, Sprache und Kultur (I:IMSK)  
Lehrstuhl für Informationswissenschaft

Modulprüfung Computational Intelligence

Modul: INF-BA-M09

SS 2021

Leitung: Stefan Kerscher/ Prof. Dr. Bernd Ludwig

**Trainieren eines neuronalen Netzes zur Prädiktion der Trajektorien von Fußgängern**

Pascal Strobel

Matrikelnummer: 2106133

6. Semester Medieninformatik/ Informationswissenschaft

E-Mail: pascal.strobel@stud.uni-regensburg.de

Abgegeben am 15.09.2021

Inhalt

[1 Einleitung 3](#_Toc82019692)

[2 Verwandte Arbeiten 4](#_Toc82019693)

[3 Methodik 4](#_Toc82019694)

[3.1 Problemstellung 5](#_Toc82019695)

[3.2 Datenvorverarbeitung 6](#_Toc82019696)

[3.2.1 Normalisieren der Daten 6](#_Toc82019697)

[3.2.2 Aufteilen der Daten in Trainings-, Validierungs- und Testdaten 7](#_Toc82019698)

[3.2.3 Standardisieren der Daten 7](#_Toc82019699)

[3.2.4 Formatieren der Daten 8](#_Toc82019700)

[3.3 Netzwerkarchitektur 8](#_Toc82019701)

[4 Evaluation 9](#_Toc82019702)

[4.1 Metriken 9](#_Toc82019703)

[4.2 Resultate 10](#_Toc82019704)

[4.3 Fälle des Scheiterns des Ansatzes 11](#_Toc82019705)

[5 Diskussion 14](#_Toc82019706)

[6 Fazit 15](#_Toc82019707)

[Literaturverzeichnis 16](#_Toc82019708)

[Erklärung zur Urheberschaft 17](#_Toc82019709)

# Einleitung

Die Vorhersage der Trajektorien von Fußgängern gewinnt mit der ansteigenden Verbreitung von autonomem Fahren und sozialen Robotern, die den Menschen am Arbeitsplatz [Socially embedded learning of the], im Krankenhaus [HelpMate], im Museum [Experiences with an interactive museum], aber auch in den eigenen vier Wänden [Care-o-bot] unterstützen sollen, zunehmend an Bedeutung. Durch präzise Vorhersagen des Verhaltens von Fußgängern in der näheren Umgebung können mögliche Kollisionen frühzeitig erkannt und verhindert werden.

In dieser Arbeit wird ein mithilfe eines tiefen neuronalen Netzes ein datengesteuerter Lösungsansatz vorgestellt, um nur mithilfe vergangener Positionsdaten die zukünftige Trajektorie eines einzelnen Fußgängers vorherzusagen. Derartig datengesteuerte Ansätze für dieses Themengebiet haben sich vor allem in den letzten Jahren durch vielversprechende Resultate etabliert.

Die Arbeit gibt zunächst einen Überblick über den aktuellen Forschungsstand. Anschließend wird der Lösungsansatz vorgestellt. Zuerst erfolgt die Definition der zugrundeliegenden Problemstellung. werden die angewandten Datenvorverarbeitungstechniken fällt die Erläuterung der

Inhaltlicher Aufbau

* Zunächst wird die eigentliche Problemstellung erläutert
* Dann wird der vorgeschlagene Lösungsansatz vorgestellt, darunter Datenvorverarbeitungsschritte, die Wahl des Neuronalen Netztes (und seine Parameter), sowie benötigte mathematischen Konzepte dahinter
* Anschließend erfolgt eine Evaluation des vorgestellten Ansatzes anhand der zwei in der Literatur gängigen Metriken ADE und FDE und es wird aufgezeigt, in welchen Fällen von Fußgängerverhalten der Ansatz versagt
* Im Anschluss werden die Ergebnisse diskutiert und Limitierungen, sowie mögliche Verbesserungsvorschläge aufgezeigt.
* Ein kurzes Fazit über den vorgestellten Ansatz rundet die Arbeit ab.

# Verwandte Arbeiten

Die „Social Force Model“ von [Social Force Model] bildet die Grundlage für physikbasierte Ansätze zur Prädiktion von Fußgänger Trajektorien. Es basiert auf der Theorie, dass Fußgänger externen Kräften ausgesetzt sind und deren Verhalten somit mathematisch beschrieben und vorhergesagt werden kann. Ein weiterentwickeltes physik-basiertes Modell zur Prädiktion von Fußgängerverhalten ist das von [BRVO] vorgestellte BRVO, welches eine bereits bestehende Fußgänger-Simulationsmethode (RVO) mit Online-Lernen kombiniert, um so eine individualisierte Vorhersage für jeden Fußgänger zu treffen.

Durch das in den letzten Jahren starke Wachstum an Popularität von künstlichen neuronalen Netzen, aber auch durch die laut [Pedestrian Trajectory Prediction with CNN] gegebene Limitierung von physik-basierten Ansätzen, dass diese aufgrund handgefertigter Funktionen nur eine Teilmenge aller möglichen Verhaltensweisen darstellen können, sind vor allem Deep Learning Ansätze zur Prädiktion von Fußgänger-Trajektorien stark in der Literatur vertreten.

Rekurrente neuronale Netze (RNN) bilden mit ihrem internen Gedächtnis eine sehr leistungsstarke Kategorie von künstlichen neuronalen Netzen und eignen sich besonders gut für die Arbeit mit Sequenzdaten, beispielsweise bei der automatischen Spracherkennung [end-to-end], maschinellen Übersetzung [Neural machine translation] oder der Klassifikation von Bildern und Videos [Look and Think twice].

Da die Aufgabe der Trajektorien-Prädiktion auch als Zeitreihenproblem angesehen werden kann, ist auch in dieser Domäne die Verwendung von RNNs weit verbreitet, speziell die der „Long Short-Term Memory“ Zellen, vorgestellt von [LSTM]. Beispiele dafür sind die Arbeiten von [Social LSTM] [Pedestrian Trajectory Prediction in extremely crowded scenarios] [Dynamic and static context-aware LSTM] [A data driven model for Interaction]. Viele der Arbeiten zum Thema Trajektorien-Prädiktion von Fußgängern beziehen neben den vergangen Positionsdaten eines Fußgängers, zusätzliche Kontextinformationen, wie beispielsweise die Trajektorien anderer Fußgänger [Social LSTM] [Pedestrian Trajectory Prediction in extremely crowded scenarios] [Dynamic and static context-aware LSTM] oder räumliche Informationen [Context-Aware Trajectory Prediction] [A data driven model for Interaction] mit in die Vorhersage ein. In dieser Arbeit ist das nicht der Fall, die Vorhersage der zukünftigen Bewegungsbahn erfolgt nur mithilfe bekannter, vergangener Positionen.

[An Empirical Evaluation of Generic] zeigen, dass faltende neuronale Netze kanonische rekurrente Architekturen in einer breiten Anzahl von Aufgaben übertreffen und folgern das nötige Überdenken der Assoziation von Sequenzmodellierung und RNNs. Derartige Architekturen sind auch bei der Vorhersage von menschlichen Trajektorien aktuell nur spärlich vorzufinden, bestärken aber die obige Schlussfolgerung und liefern zuversichtliche Resultate [Pedestrian Trajectory Prediction with Convolutional Neural Networks | Convolutional Neural Network for Trajectory Prediction].

# Methodik

In diesem Abschnitt wird zunächst die Problemstellung dargelegt und die informationswissenschaftliche Fragestellung abgeleitet. Anschließend wird ein möglicher Lösungsansatz präsentiert, welcher neben den verwendeten Vorverarbeitungstechniken der Daten und der Wahl des trainierten neuronalen Netzes und seiner Hyperparameter, auch die zugehörigen mathematischen Grundlagen erläutert.

## Problemstellung

Die Prädiktion der Trajektorien von Fußgängern entspricht der Vorhersage des zukünftigen Bewegungspfades basierend auf einer bestimmten Anzahl zuvor beobachteter Positionen. Die informationswissenschaftliche Fragestellung, die sich daraus ergibt, und für die diese Arbeit einen Lösungsansatz liefert, lässt sich demnach folgendermaßen formulieren:

„Wie kann anhand bekannter, vorheriger Positionen die zukünftige Trajektorie eines Fußgängers bestimmt werden?“

Zur Beantwortung der Frage soll auf Basis eines gegebenen Datensatzes, welcher Daten von sich bewegenden Fußgängern enthält, ein neuronales Netz trainiert und eine Regression durchgeführt werden. Jeder Datenpunkt im Datensatz besteht dabei aus einer Zeitreihe mit insgesamt 20 Zeitschritten und den dazugehörigen Werten für die aktuelle x- und y-Koordinate (in Metern) des Fußgängers. Die zeitlichen Intervalle zwischen zwei Zeitpunkten innerhalb eines Datenpunktes betragen durchgehend 400ms. Neben den Koordinaten sind keine weiteren kontextuellen Informationen über den Fußgänger oder seine Umgebung bekannt.

Das trainierte neuronale Netz soll anschließend dazu genutzt werden, mithilfe der Informationen aus den ersten acht Zeitpunkten eines Datenpunktes, die zukünftigen zwölf Positionen des Fußgängers vorherzusagen.

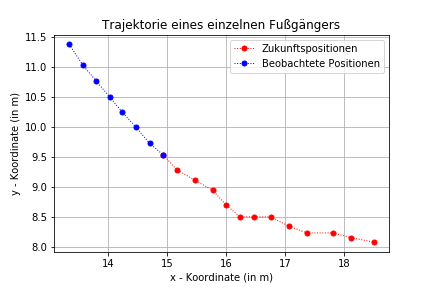


Abbildung 1: Beispiel-Trajektorie eines Fußgängers. Mithilfe der ersten acht beobachteten Positionen (blau) sollen die zukünftigen zwölf (rot) vorhergesagt werden (eigene Darstellung)

Alle Grafiken mit [Matplotlib] erstellt 🡪 Fussnote machen

Diese acht-zu-zwölf-Verteilung ist gängig in der Literatur und beispielsweise in den Arbeiten von [Social LSTM: Human Trajectory Prediction in Crowded Spaces | Pedestrian Trajectory Prediction with Convolutional Neural Networks | Convolutional Neural Network for Trajectory Prediction] zu finden. Die zugrunde liegende Aufgabe wird in Abb. 1 zusätzlich visualisiert.

## Datenvorverarbeitung

### Normalisieren der Daten

Der ursprüngliche Datensatz enthält die Positionen der Fußgänger als absolute Koordinaten. Da allerdings kein Ursprung für diese Koordinaten festgelegt ist, können die Bewegungsdaten unterschiedlicher Fußgänger absolut weit auseinander liegen (Fußgänger 1 z. B. im Bereich 1 bis 10m, Fußgänger 2 im Bereich 90 bis 100m).

[Pedestrian Trajectory Prediction with Convolutional Neural Networks] identifizieren neben dem naiven Ansatz, die absoluten Koordinaten ohne weitere Normalisierung zu verwenden, drei mögliche Normalisierungstechniken:

* Koordinaten haben den Ursprung im ersten beobachteten Zeitpunkt
* Koordinaten haben den Ursprung im letzten beobachteten Zeitpunkt
* Relative Koordinaten

Alle drei Ansätze führen dabei zu einer Performancesteigerung des in der Arbeit vorgestellten, faltenden neuronalen Netzes für unterschiedliche Datensätze, was die Notwendigkeit der Normalisierung der Daten nochmals betont.

In dieser Arbeit werden die Daten gemäß dem zuletzt aufgeführten Stichpunkt normalisiert. Das bedeutet, das das später trainierte neuronale Netz nur mit den relativen Änderungen der x- und y-Koordinaten zwischen zwei Zeitpunkten arbeitet, welche im weiteren Verlauf der Arbeit als Delta-x (dx) und Delta-y (dy) bezeichnet werden. Zur Bestimmung der Delta-Werte wurde für jeden Zeitpunkt eines Datenpunktes für alle Datenpunkte im Datensatz die Differenz der x- und y-Koordinaten vom aktuellen Zeitpunkt t und zum vorherigen Zeitpunkt t-1 berechnet. formal dargestellt (evtl. Formel mit einfügen):

Das führt zu einer nötigen Spezialisierung der Anforderungen an das neuronale Netz: Aus den ersten acht Zeitschritten eines Datenpunktes werden je sieben dx- und dy-Werte berechnet. Das neuronale Netz soll demnach Zeitreihen der Länge sieben und den zwei Features dx und dy entgegennehmen und dafür die nächsten zwölf dx- und dy-Werte vorhersagen. Fussnote: Die Umwandlung in die eigentlichen Koordinaten erfolgen ganz einfach durch das addieren der jeweiligen vorhergesagten Delta-Werte auf die vorherigen Koordinaten

### Aufteilen der Daten in Trainings-, Validierungs- und Testdaten

Die normalisierten Datenpunkte wurden nach dem Zufallsprinzip in Trainings- (80%), Validierungs- und Testdaten (je 10%) aufgeteilt.

Zur Vermeidung von Data-Leakage erfolgte diese Aufsplittung vor der Standardisierung der Daten.

### Standardisieren der Daten

Um sicherzustellen, dass die Wertebereiche der zwei Merkmale dx und dy nicht zu unterschiedlich sind, werden diese mithilfe der z-Transformation standardisiert, sodass für jedes Merkmal ein Erwartungswert von 0 und eine Standardabweichung von 1 vorliegt. Der -Score des Samples lässt sich mit folgender der Formel berechnen:

ist dabei das arithmetische Mittel und die Standardabweichung der Trainingssamples.

Eine Verlagerung der Mittelwerte der Inputvariablen gegen null, wie es bei der z-Transformation der Fall ist, wird auch von [EfficientBackprop] empfohlen, da jede Verschiebung des durchschnittlichen Inputs weg von null die Updates der Gewichte in eine bestimmte Richtung verzerrt und damit den Lernvorgang des neuronalen Netzes verlangsamt.

Die Standardisierung der Trainingsdaten, sowie die später folgende Umkehrung der z-Scores erfolgte mithilfe der StandardScaler- Klasse der Scikit-learn Bibliothek. (evtl. als Fußnote)

### Formatieren der Daten

Zur Bestimmung der nächsten zwölf Positionen eines Fußgängers wird in dieser Arbeit die Strategie der rekursiv-mehrstufigen Vorhersage angewandt. Das bedeutet, das NN prädiziert basierend auf sieben dx und dy Werten immer nur den nächsten dx und dy Wert, statt alle zwölf auf einmal. Die neugewonnenen Informationen durch das NN werden anschließend an die vorherigen sechs Inputdaten angehängt (der erste Wert fällt weg, da das NN als Input eine Zeitreihe der Länge sieben erwartet), was wiederum den Input für die nächste Prädiktion bildet. Dieser Vorgang wird zwölfmal wiederholt.

Damit das NN für diese Aufgabe trainiert werden konnte, mussten die Trainingsdaten entsprechend angepasst werden. Jeder Datenpunkt des Trainingsdatensatzes wurde daher in zwölf Trainingssamples zerlegt. Immer sieben aufeinanderfolgende dx und dy Werte bilden dabei den Input, und der darauffolgende dx und dy Wert den gewünschten Output/ das Orakel eines Trainingssamples.

## Netzwerkarchitektur

Verwendetes Netzwerk: RNN, da RNN sich besonders gut für Daten im Zeitreihenformat eignet, allerdings hat RNN noch einige Probleme

* Deshalb LSTM
* LSTM hat sich auch in der Literatur etabliert
* Fußnote: Aber auch in kurzlich veröffentlichten Artikeln schneiden CNN sehr gut ab!
* Was ist LSTM/ CuDNNLSTM
* Was ist eine Dense Layer
* Was sind units
* Welche methoden wurden gegen Overfitting angewendet? 🡪 L1-Regularisierung 🡪 wie funktioniert das?
* Learning rate
* Epochen
* Adam optimizer
* Loss funktion

# Evaluation

## Metriken

Ebenso wie in verwandter Literatur von bspw. [Pedestrian Trajectory Prediction with Convolutional Neural Networks | Convolutional Neural Network for Trajectory Prediction] werden zur Auswertung der Performance die Metriken „Average Displacement Error“ (ADE), erstmalig eingeführt von [You’ll Never Walk Alone: Modeling Social Behavior for Multi-target Tracking], und „Final Displacement Error“ (FDE) eingesetzt.

Der ADE eines Fußgängers ist die durchschnittliche euklidische Distanz zwischen den prädiktierten Positionen und der Ground-Truth für alle vorhergesagten Zeitschritte. Der ADE lässt sich mit folgender Formel berechnen:

ist der Zeitschritt der letzten Beobachtung des Fußgängers (in unserem Fall gilt = 8), der final zu vorhersagende Zeitschritt eines Datenpunktes (in unserem Fall gilt = 20). ist die vorhergesagte und die tatsächliche Position zum Zeitpunkt t. repräsentiert die euklidische Distanz.

Der FDE eines Fußgängers ist die euklidische Distanz zwischen der prädizierten Position und der zugehörigen Ground-Truth zum finalen Zeitpunkt und lässt sich folgendermaßen berechnen:

## Resultate

Die Performance des trainierten RNN wurde anhand des zu Beginn abgespaltenen Testdatensatzes (10% der Gesamtdaten), welcher die Trajektorien von insgesamt 540 Fußgängern umfasst, ausgewertet, indem für jeden Datenpunkt sowohl der ADE als auch der FDE bestimmt wurde. Damit eine gesammelte Aussage für die Performance des Systems gemacht werden kann, wurde der Durchschnitt der jeweiligen Metrik von alle 540 Datenpunkte berechnet, was zu folgendem Ergebnis führt:

Das RNN erzielt für die Testdaten einen durchschnittlichen ADE von 0,763 und einen durchschnittlichen FDE von 1,561.

Da x- und y-Koordinaten der Fußgänger in Metern angegeben sind, kann der ADE auch als der durchschnittliche Abstand in Metern zwischen einer vom System prädiktierten Zukunftsposition und der Grundwahrheit angesehen werden. Äquivalent dazu beschreibt der FDE den durchschnittlichen Abstand in Metern zwischen der prädiktierten und der tatsächlichen Position des Fußgängers zum finalen Zeitpunkt.

## Fälle des Scheiterns des Ansatzes

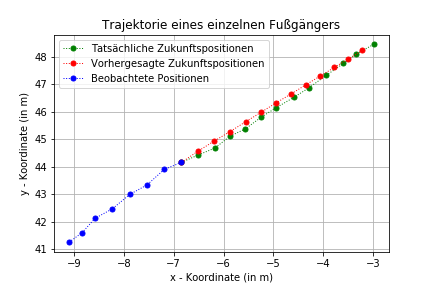


Abbildung 2: Präzise Vorhersagen durch triviales Fußgängerverhalten (eigene Darstellung)

Nur für wenige Fußgänger ist die Vorhersage der zukünftigen Trajektorie so trivial wie in Abb. 1. Abrupte Richtungswechsel oder Geschwindigkeitsänderungen sorgen für ein oft unberechenbares Verhalten von Fußgängern und damit zu Einbußen in der Performance. Unzuverlässige Prädiktionen des vorgestellten Systems sind dabei vor allem bei den folgenden vier Fällen (erstmal drei) zu beobachten.

1. Fußgänger ändert plötzlich Richtung oder Geschwindigkeit, ohne einen Hinweis darauf:

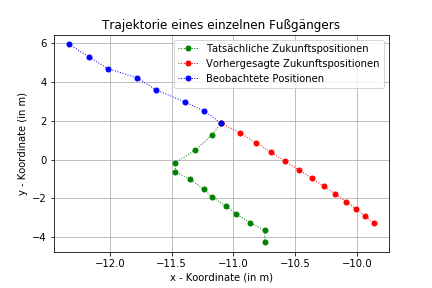


Abbildung 3: Überraschende Richtungsänderung (eigene Darstellung)

Die beobachtete Trajektorie des Fußgängers in Abb. 2 gibt keinerlei Hinweise auf eine mögliche Richtungsänderung. Die bestmögliche Vorhersage, die vom System (aber auch vom Menschen) getroffen werden kann, ist das Weiterführen dieser Bewegungsbahn. Damit derartige Bewegungsmuster besser prädiziert werden können, sind zusätzliche Kontext-Informationen (bspw. Straßenverlauf, Aufeinandertreffen mit anderen Fußgängern o.ä.) über den Fußgänger notwendig.

1. Fußgänger ändert mehrfach seine Richtung oder Geschwindigkeit:

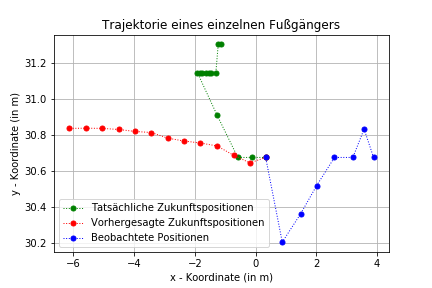


Abbildung 4: Mehrfache Richtungs- und Geschwindigkeitsänderungen (eigene Darstellung)

Nicht nur plötzliche, unerwartete, sondern auch häufige Richtungs- und Geschwindigkeitsänderungen in den beobachteten Positionen, führen zu Schwierigkeiten bei der Vorhersage, wie in Abb. 3 ersichtlich. Ein richtiges Muster ist hierbei auch für den Menschen schwer erkennbar. Wie auch schon im ersten Fall, sind zusätzliche Kontext-Informationen nötig, die eine zuverlässigere Prädiktion erlauben.

1. Fußgänger ändert seine Position nicht:

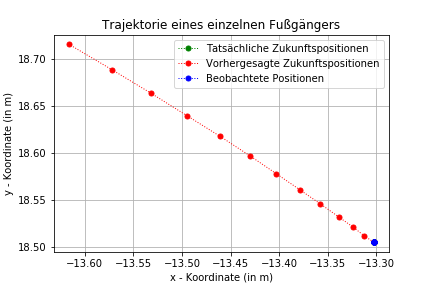


Abbildung 5: Keine Positionsänderung (eigene Darstellung)

Ebenfalls schlechte Vorhersagen des Systems sind zu beobachten, wenn sich der Fußgänger nicht bewegt, wie es bspw. in Abb. 4 der Fall ist. In derartigen Fällen scheint das System nicht gelernt zu haben, keinerlei Bewegung in x- und y- Richtung vorherzusagen. Stattdessen werden zukünftige Werte für dx und dy prädiziert, die leicht von Null abweichen. Durch das rekursive Vorgehen bei der Bestimmung der nächsten zwölf dx- und dy-Werte summiert sich dann der Fehler und führt zu zunehmenden Abweichungen zwischen Prädiktionen und Ground-Truth.

1. Fußgänger bewegt sich nach einem komplexeren Muster:

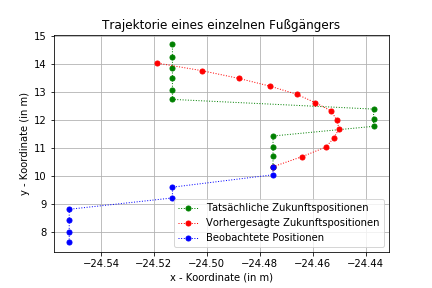


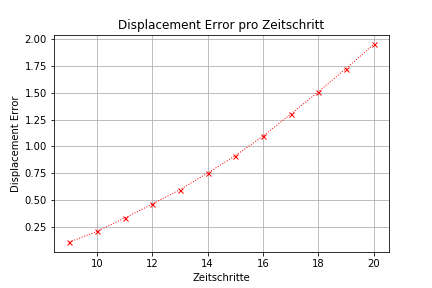
Abbildung 6: Komplexes Bewegungsmuster (eigene Darstellung)

* Bessere Abbildung finden, bei der der Mensch das Muster besser erkennen kann.
* Erst mal optional

Bei komplexeren Trajektorien

# Diskussion

Die

* Iteratives / Rekursives Predikten hat größeren Fehler zur Folge 🡪 siehe Grafik 🡪 besser wäre direkte Prediktion von 12 nächsten Pos
* Multi vs. Sequential Output. Trajectory prediction sequentially point-by-point performs poorly due to error propogation to future time-steps (trajectory curves off). Our multi-output model tends to be more resistant to such error accumulation. (aus dem zweiten Trajectory prediction paper with cnn) 
* Nur begrenzt Trainingsdaten
* Als Lösung: Data Augmentation (Rotieren, etc.)
* Hyperparameter wurden manuell optimiert / hyperparameter Tunin
* Nicht alle unterschiedlichen Kombinationen konnten getestet werden (z.B. wurde nur CudNNLSTM verwendet, da schneller, allerdings kann die activation Funktion nicht geändert werden, tanh steht sicher)
* Modell auf andere Datensätze (Trajnet etc.) anwenden, um mit bereits etablierten Ansätzen vergleichen zu können

🡪 two publicly available datasets: ETH [49], and UCY [39]. 🡪 social LSTM

* Auch komplett anderere Ansätze funktionieren, z.B. das Verwenden von CNN (siehe Paper) 🡪 vielleicht eher im Fazit

# Fazit

Diese Arbeit präsentiert einen möglichen Lösungsansatz zur Prädiktion von Fußgänger-Trajektorien mithilfe eines rekurrenten neuronalen Netzes. Dabei werden alle Datenvorverarbeitungsschritte, sowie das Training des neuronalen Netzes Der Ansatz liefert zufriedenstellende Ergebnisse für triviale Bewegungsbahnen

Trajektorie Predicition ist ein schwerer Task, was in diesem Paper auch nochmal klar wurde. Gerade aber für Anwendungen wie autonomes Fahren oder selbstfahrende Roboter am Arbeitsplatz, ist es enorm wichtig, sehr präzise Vorhersagen zu treffen, um Kollisionen mit Fußgängern zu verhindern.

Für triviale Vorhersagen ist das System zuverlässig, allerdings für die oben genannten Fehlerfälle nicht (da Fußgänger bewegung allerdings meistens unberechnbar sind, sind Änderungen am System, die eine präzisere Prädiktion, auch für nicht-triviale Cases, unbedingt notwendig)

Recursive Multi-Step Forecast

Literaturverzeichnis

Erklärung zur Urheberschaft

Wir haben die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie alle Zitate und Übernahmen von fremden Aussagen kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt.

Die vorgelegten Druckexemplare und die vorgelegte digitale Version sind identisch.



Regensburg, 14.09.2021

|  |  |  |
| --- | --- | --- |
| Ort, Datum |  | Unterschrift |